Response to RC2
https://doi.org/10.5194/egusphere-2024-3255-RC2

The authors expand existing methodologies for the analysis of spatiotemporal relationships between flood-related topics in social media posts, flood, and basin characteristics during major flooding events. The proposed analysis is interesting and potentially useful to flood-response authorities. The manuscript is well written.

I have some recommendations to enhance clarity and facilitate reproducibility across other case-study regions, before publication:

1. At the end of Sections 2.2.1 and 2.2.2, I would include a table summarizing meteorological, flood, and watershed data used in the study.

Thank you for this suggestion. A table summarising the total watershed area, precipitations, flooded areas, along with the average population density, and number of geotagged tweets is now provided at the end of the data section. (L159-160)

2. Lines 177-180: briefly describe the zonal statistical approach and the coverage fraction method used to assign precipitation values from the raster dataset to the watersheds.

Here is the precision we made:

"Daily precipitation data were aggregated at the watershed level using a zonal statistics method. We employed a coverage fraction technique (weighted sum) to summarize the raster precipitation values within each watershed polygon. We choose the weighted sum method that multiplies the precipitation amount of each grid cell by the fraction of the cell contained within the watershed, thereby refining sub-estimates of total precipitations per watershed" (L184-190)

3. Lines 167-170: river catchment areas most affected by flooding are identified as those with more than 100 mm of precipitation. How do the authors consider any major effects in downstream locations that were not directly involved by high precipitation?

This important information was accounted for in our study using the extend of flooded zones from the Mapping Portal of the Copernicus EMS (Wania et al., 2021). A complete description of this layer is described in Section 2.2.1 (L122-128). This variable is then analysed in detail based on Figure 7 (section 3.5). Although many downstream areas did not receive heavy rainfall, they experienced widespread river overflooding which triggered the generation of a significant number of Tweets.

4. Lines 178-180: knowing the packages used to perform the analysis may not be sufficient to reproduce the analysis. I would invite the authors to share a code streamlining (at least parts of) the methodological steps. That would help local authorities and stakeholders take advantage of geo-social data during flood hazard management. I suggest the authors include a code availability statement in their article. If the authors used code from other sources, then they could include the full list of those sources in that statement.

Thank you for your suggestion. We do agree that a streamlined version of the major methodological steps we used in our analysis can help authorities and stakeholders to employ our results in future disaster situations. Therefore, we added the following code availability statement (L625):

“The code to reproduce the Tweet translation, topic modelling and the topic coherence analysis can be found in the corresponding GitHub repository: https://github.com/DorianZGIS/Tracing-online-flood-conversations-across-borders.git.“

5. Lines 192-198: briefly explain what “embedding” and “vectorization” mean in the context of natural language processing

Thank you for your suggestion. On to what we mean with embeddings in this context:

“First, it converted the individual Tweet texts into numerical representations by creating embeddings using a BERT-based algorithm (in our case, multi-qa-distilbert-cos-v1), which maps words into a vector space designed to preserve semantic relationships.” (L203-206)

We also agree that the term vectorisation needs some more clarification in this context. Therefore, we enhanced the explanation of the fourth step in our analysis:

“In the fourth step, we used a CountVectorizer from scikit-learn to transform a list of stop words into word vectors, explicitly excluding them to allow for more meaningful topic formation.” (L211-2012)

6. Line 202: is the number of topics (30, in the specific case) a parameter of the k-means clustering algorithm? E.g., a predetermined number of clusters that the algorithm is asked to find.

Thank you for your comment. Yes, the number of topics is a parameter of the k-means clustering algorithm determining the number of clusters to find. We further clarified this in our description of the individual topic modelling steps:

“We achieved the best results in terms of topic coherence by utilising a K-Means clustering algorithm, where the number of topics identified corresponds to the predetermined number of clusters.” (L209-211).

7. Lines 224-227: in those cases where more than one topic was equally dominant in some days, how did the authors decide what topic to retain and what others to discard? How often does this happen? Is there any risk of introducing subjectivity?

We implemented what can be referred as a heuristic approach, i.e. a practical method or rule-of-thumb strategy that we developed for tracking the evolution of online conversions and detect their dominance over time.

While an exact solution such as weighted averages of topics for instance would have complicated interpretations, this allowed to provide a simple but robust method able to select one dominant topic per watershed on given day.

Instead of computing the relative importance of each topic, our method assigns the topic name that is the most discussed per watershed on a given day based on two main heuristic rules:

1) The selection of the most discussed topic

2) The removal of topics with the same number of maximum occurrences.

This approach was clarified throughout the section (see 2.3.3.).

Further, we discuss the risk in the discussion section (see 4.2.4) by providing two additional tables (see Table S2 and Table S3) supporting the following points:

-The less frequently discussed topics were primarily affected by the filtering process. For instance, 61% of topics mentioned only once on a given day and watershed were discarded. In contrast, none of the most frequently discussed topics (i.e. those occurring between 11 and 35 times) were removed from the analysis. Visually, we could also observe that those small topics corresponded to more peripheral places located outside the most impacted areas, hence carrying some noise that was removed thanks to this this method.

-Topic-wise, the filtering process impacted each topic to a relatively similar degree, with the proportion of discarded topics ranging from 40% to 70%, except for Roads Blocked, for which only 24.8% of occurrences were filtered out. We found that this topic was a rather atypic topic generated from the *Touring Mobilis* twitter account. It described the different traffic problems occurring throughout our study area, which explains why it was less often in conflict with other topics and was thus kept in the analysis.

8. Fig. 3: I suggest mentioning in the caption the HydroBASIN watershed level considered.

Caption change made (see Fig. 3).

9. Fig. 4: for each daily bar, what are the empty portions on top of the shaded portions? In other words, what are daily percentages calculated on? Is 100% the total amount of Tweets, including those unrelated to flooding?

Indeed, the remaining percentages corresponds to Tweets that are not related to flooding. We added this precision in the figure caption (see Figure 4)

10. Fig. 7 is very useful to outline the spatial distribution of social-media topics. However, some additional clarifications are necessary. Does each distribution really represent the frequency of dominant topics (as stated in the caption)? Or else, since there is a distribution for all considered topics, what the figure really shows is the frequency (in space) of each topic one at a time? Given that some topics were more frequent than others, is it the case that different distributions are associated with different overall numbers of topic occurrences in space? Also, what about the temporal variability? Does the figure show the cumulative values of topic occurrences at any basin locations, cumulated over time? Or something else? The authors should include a more detailed explanation to clarify these aspects.

Indeed, each distribution does represent the frequency of a given dominant topic.

We did not analyse the temporal variability in this plot. This is already provided with figure 3. We aggregated over time the number of times a topic was dominant across the study area.

To analyse the spatial variability, we classify watersheds based on their attribute values by creating 100 quantiles and analysed the number of times a given topic dominates the conversations over the period from 7 to 27 July.

Additional clarifications were provided in L275-284.

11. Fig. 7 is very effective in showing the spatial variability in flood-related topics discussed in social media. However, as it is now, it does not consider how these distributions may change across different macro regions. Given the emphasis that the authors give to the trans-boundary character of their study (e.g., lines 17, 24, 38), I think it would be interesting to include three more figures like Fig. 7 but referred to the

individual Escaut, Meuse, and Rhine river basins, to see if any significant differences emerge.

Thank you for this suggestion. The transboundary nature of our study is indeed a crucial aspect that we aim to develop further.

In the early stages of this research, we generated separate plots for each watershed to analyse regional differences more effectively. Our analysis revealed distinct variations between watersheds, driven by region-specific factors such as the location of urban centres and the distribution of the most impacted areas along the river profile.

However, we also observed that the number of dominant topics identified was sometimes too limited to draw robust conclusions, necessitating further investigation. In some cases, the number of dominant topics was too low for a meaningful interpretation—for example, only a single dominant topic was identified for the 'Rhine flood' in the Escaut River. This limited number of occurrences made the distribution plots for certain watersheds difficult to interpret.

By contrast, the aggregated ridgeline plot for the entire study area contained enough data to provide meaningful insights and draws more robust conclusions. Additionally, it allowed for a synthesis of the observations made at the main basin level, offering a comprehensive summary of the overall trends observed across our study area.

To supplement the analysis provided in the manuscript, we have now included a detailed table (see Table S2) with the key figures discussed above and a figure showing the distribution of each topic per main basin (see Figure S4).

12. Lines 503-504: Under what label were the "Damage"-centered tweets clustered at those iterations where the "Damage" topic did not emerge? Were they clustered with the Rhine and Meuse flood topics? Would the results change remarkably if the "Damage" topic were not considered?

Thank you for the opportunity to clarify this statement.

We adjusted section: 4.2.3 accordingly to clarify this point.

"An additional consideration in our analysis was the inherent variability of the semantic modelling algorithm (BERTopic) (Grootendorst, 2022b), which is not entirely deterministic and depends on randomness in identifying topic clusters. To mitigate this issue, we ran the algorithm 20 times to assess the stability of the topics, distinguishing between stable and unstable clusters. We found that the keyword sets defining the Topic 26 and Topic 10 of the *Damages* topic only occurred across 20% and 30% of the iterations when allowing for a maximal difference of 17%. Similarly, topic 18, which was aggregated into the *Compassion* topic, was identified in only five iterations (25%), and the *Help to Victims* topic was stable across nine iterations (45%).

This is due to the fact that we applied a highly restrictive maximal difference threshold of just 17% between topics across iterations, which, in some cases, corresponded to a difference of fewer than eight characters. Thus, this does not imply that other iterations were missing topics related to damages, help to victims, or compassion. Rather, it means that the defining keywords for these topics changed by more than 17%, exceeding our threshold and resulting in their classification as distinct topics when comparing across iterations. This variability in keywords needs to be considered when interpreting these less stable topics. However, future studies could enhance topic stability analysis by incorporating ensemble approaches, combining results from multiple iterations to form a consensus topic structure, or by exploring more sophisticated embedding-based similarity comparisons, which allow to capture the underlying meaning of the keywords."

We also added the following clarification to the section 2.3.2:

"To compare how frequently a specific topic appeared across iterations, we sought to identify a threshold for the maximum allowable difference between topics. The reasoning behind this being that the keywords defining topic A in iteration 1 can be split across several different topics in other iterations. Hence, topic A in iteration 1 could in theory be matching with several to almost all topics in iteration 2. As a result, some matches will be very weak or even mismatches, especially if we allow for characters of words to also change slightly change. Thus, to mitigate the likelihood of mismatches, we defined an upper bound threshold for how many changes are allowed between to topics of different iterations to be classified as a match. We found that a 17% difference provided such an upper bound, accounting for slight changes in defining words, while ensuring that one topic was only matched to one other topic per iteration. To assess differences between keywords defining each topic, we employed the string edit distance. Finally, we chose the topic model iteration for our subsequent analysis which exhibited the most stable topics (see Figure S2 for more details). "

Minor comments:

1. Line 123: correct "extend" to "extent"

Change made.

2. Figure 1: the part of study area within Belgian boundaries presents an abrupt jump in

the gradient of greys representing elevations.

This is because of the blue fill that has different levels of transparency. If we remove this layer, we lose the distinction between the three watersheds. A legend was added as to label the colour scheme chosen for highlighting the 3 different river basins studied.

3. Figure 1: the lower bound of the elevation color bar is negative, and equal to -179 m; what is the reference elevation associated with 0 m?

0 is the sea level. -179 represents coal mines in Germany and/or polders areas in the Netherlands. We replaced this negative value by 0 as to avoid confusion for the readers.

4. Line 101: do not use parentheses inside parentheses for the citations' years.

Thank you for your eye for detail, we corrected this: (e.g. Wang et al. 2018, Tan and Schultz 2021) and for (e.g. Kruspe et al. 2021, Zou et al. 2018).

5. Lines 194-195: explain what the acronym UMAP stands for, exactly.

Thank you for the comment. UMAP stands for Uniform Manifold Approximation and Projection and is a dimensionality reduction algorithm. We clarified this in the text. (L206)

6. Caption of Fig. 3 (line 290) says that thick blue is the color used to represent rivers most impacted by precipitation. However, this is not in agreement with the legend, which adopts a light orange instead.

Thank you for noticing. Change made.

7. Table 1: there are some words (e.g., “maas”, “venlo”, “dinant”, “nrw”, etc.) that are not in English. If they are names of towns, cities, rivers, etc., I would specify that somehow at the bottom of the table. Otherwise, some readers that are not familiar with the local geography might be left wondering whether they are untranslated words.

Thank you for this comment. We added footnotes to clarify the meaning of these words. (see Table 2).

8. Fig. 7: clarify in the caption what watershed scale was used.

Plots of Fig. 7 represents aggregated values for the entire study area with the frequency representing the number of tweets per watershed. This precision was added (see Fig. 7).

9. Line 453: “show” should be “shows” for third person singular

Thank you for noticing this error, we made the change.

10. Line 461: correct “use” to “used”

Thank you for your eye for detail, we made the change.

our eye for detail, we made the change.